Patent claims

1. Method for the determination of potentially important DNA and/or nucleic acid sequences of a species of interest (species sequences) with the following steps:

a) determination of any desired species sequences of the species of interest by biological or genetic engineering methods and storage of the species sequences in a first databank,

b) acquisition of known DNA/nucleic acid sequences of a given group of other species (biosequences), including the functional importance of these sequences, in a second databank in which the biosequences and additional information, including the functional importance of individual biosequences, are stored,

c) comparison of the already known species sequences of the species of interest with the biosequences of the given group of biosequences stored in the second databank in a homology test,

d) separating out of those biosequences of the given group which are homologous to the known species sequences above a given threshold value,

e) comparison of the biosequences from the group mentioned which remain from the second databank and have not been separated out with the species sequences determined according to step a in a second homology test,

f) storage and/or issuing of those species sequences as species sequences of potentially increased importance, homology of which with biosequences from the biosequences remaining from the group mentioned exceeds a given second threshold value, together with information on the biosequences in each case homologous thereto,

g) it being possible optionally also to carry out step e) before step c) and without prior separating out according to step d).

2. Method according to claim 1, characterized by the following further steps:

h) matching of the species sequences issued or stored in step f) in a matching, optimized according to criteria which can be predetermined, to the particular homologous biosequences and issuing and/or storage of characteristic parameters of the optimized match, such as, for example, the percentage agreement, the length of coinciding sequence sections and the optimized relative alignment.

3. Method according to claim 1, characterized by the following further steps:

i) classification of the species sequences issued or stored in step f), i.e. assignment (sorting) into particular classes of sequences by linguistic analysis of text definitions of the additional information stored on the homologous biosequences.

4. Method according to one of claims 1 to 3, characterized by the following step:

k) addition to the characteristic information of the particular homologous biosequences which is to be assigned to the potentially important species sequences by acquisition of references (links) relating to the biosequences acquired according to step f) in the second databank to at least one third databank and acquisition of the information stored in the third databank on the biosequences mentioned.

SUB A1

5. Method according to one of claims 1 to 4, characterized in that the third databank holds ready a classification organized taxonomically at least in part regions.

6. Method according to claim 5, characterized in that the third databank is the MEDLINE databank.

7. Method according to claim 5, characterized by comparison of the keywords assigned to the particular biosequences according to taxonomic criteria with a given list or file of keywords and issuing of coinciding keywords as well as the biosequences in question and the homologous species sequences or in each case an identification thereof for which keywords which coincide with the given list of keywords have been found.

8. Method according to claim 2 and one of the claims referring back to claim 2, characterized in that the comparison of a given (classified) list of keywords is carried out at least with the Medical Subject Headings of the Medline Databank.

SUB A2

9. Method according to one of claims 1 to 4, characterized in that the third databank is the UNIGENE databank.

10. Method according to claim 9, characterized in that on the basis of the EST cluster positions from UNIGENE, information on corresponding or adjacent sequence sections is acquired from GENEMAP and/or GDB.

11. Method according to claim 1 or 2, characterized in that further databanks are searched for linking members to the citation determined in the third databank, and addition of the corresponding further information or of references to the further information to the corresponding species sequences of increased importance.

12. Method according to one of claims 1 to 11, characterized in that at least the second databank is a databank accessible to the public.

13. Method according to one of claims 5 to 12, characterized in that the further databanks are chosen from the group consisting of the Unigene, Genemap and GDB (new) and OMIM, KEGG and UMLS databanks.

SUB A3

14. Method according to one of claims 1 to 13, characterized in that the addition of further information to the species sequences determined according to step f is carried out in a pipeline method, the information added being in the form of linking members to the assigned positions in further databanks.

15. Method according to one of claims 1 to 14, characterized in that the species of interest is the human species, and in that the assigned group of biosequences comprises the biosequences of invertebrate animals, mammals, primates, rodents and vertebrates, and the not yet classified new entries of the second databank.